US DOE SC ASCR

Software Effectiveness

SC GG 3.1/2.5.2 Improve Computational Science Capabilities

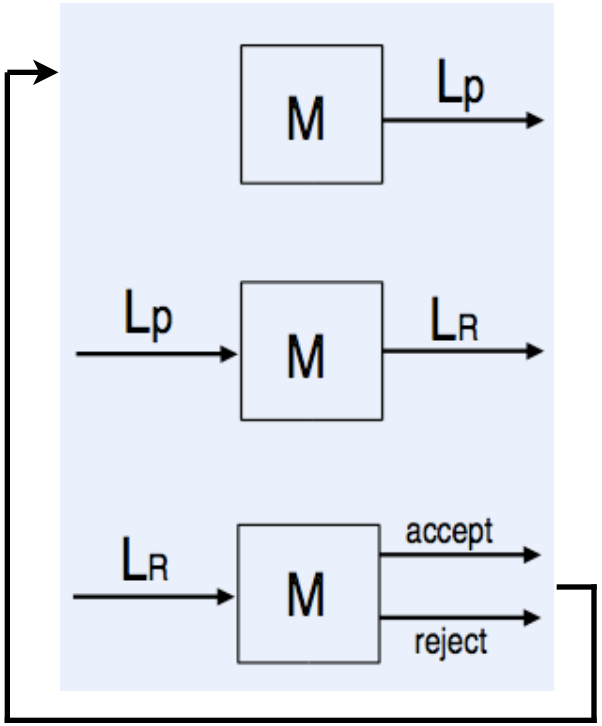
GOAL

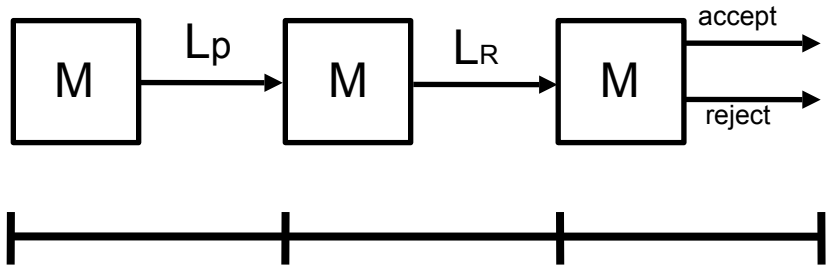
(SC GG 3.1/2.5.2) Improve computational science capabilities, defined as the average annual percentage increase in the computational effectiveness (either by simulating the same problem in less time or simulating a larger problem in the same time) of a subset of application codes. *Efficiency measure: X% (FY09, x=100)*

US OMB PART Annual Goal with Quarterly Updates

Metric: the distance between two points in some topological space

- PROBLEMS
- ALGORITHMS
- MACHINES
  - COMPLEXITY





Measured time for machine M to generate the language of the problem plus time to generate the language of the result plus the time to accept or reject the language of the result.

- Asking questions, solving problems is recursive process
- Accepting a result means a related set of conditions is satisfied

$$S = S_1 \wedge S_2 \wedge \ldots \wedge S_n$$

“simulating the same problem in less time”

- Algorithm, machine strong scaling :
  - Q4 problem := Q2 problem
  - Q4 algorithm := Q2 algorithm
  - Q4 machine ~ k * Q2 machine
  - Q4 time ~ 1/k * Q2 time

- Algorithm enhancements, performance optimizations:
  - Q4 problem := Q2 problem
  - Q4 algorithm ~ enhanced Q2 algorithm
  - Q4 machine := Q2 machine
  - Q4 time ~ 1/k * Q2 time

*Could consider other variations:
algorithm and machine are varied to achieve reduction of compute time

“simulating a larger problem in same time”

- Algorithm, machine weak scaling (defined as 100%):
  - Q4 problem ~ k * Q2 problem
  - Q4 algorithm := Q2 algorithm
  - Q4 machine ~ k * Q2 machine
  - Q4 time := Q2 time

- Algorithm enhancements, performance optimizations:
  - Q4 problem ~ k * Q2 problem
  - Q4 algorithm ~ enhanced Q2 algorithm
  - Q4 machine := Q2 machine
  - Q4 time := Q2 time

*Could consider other variations:
problem, algorithm and the machine are varied to achieve fixed time assertion

Computational Efficiency

- Total elapsed time to execute a problem instance with a specific software instance (algorithm) on a machine instance
- Parallel
  - $e(n,p) := T_{seq}(n) / (p * T(n,p))$

EXAMPLE: Efficiency measure is x% and <100%.

Enhanced Efficiency Assertion ~

$(T_{Q2} - T_{Q4}) / T_{Q2} = (kT_{Q4} - T_{Q4}) / kT_{Q4} = (k - 1) / k = x \ (k > 1)$

E.G. If x% = 50% , then k = 2. Speedup factor of two is required!

Benchmark Trends (FY04 - FY09)

Cray	X1
	X1E
	XT3
	XT4
	XT5
IBM	SP Power3
	P690
	Power5
	BG/L
SGI	Altix
HP Itanium-2	
QCDOC	

FY06: ~ 211,888 cpu-hours
FY07: ~ 314,459 cpu-hours
FY08: ~ 2,718,788 cpu-hours
FY09: ~ 16,807,139 cpu-hours*

climate research	3
condensed matter	2
fusion	5
high energy physics	2
nuclear	1
subsurface modeling	1
astrophysics	2
combustion chemistry	4
bioinformatics	1
math, data analytics	2
molecular dynamics	1
Total	24

application software	application metric	target platform	problem	joule result
(08)DCA++	time/disorder configuration	Cray XT5 31272pes, 23791s	256 disorder configs, 150nts in 2D Hubbard Model - impact on Tc	weak(k=4)
(08)GYRO	timesteps / second / process	Cray XT5 24576pes, 152.75s	improve the electron to ion mass ratio (mu=40, 20ts) in magnetically confined tokamak plasma	weak(k>5)
(08)PFLOTRAN	time/dof/PE	Cray XT5 8000pes, 2958.36s	764 x 1414 x 120 grid resolution of reactive transport in Hanford 300	weak(k=2)
(07)CHIMERA	Compute time / subcycled hydrodynamic step	Cray XT3 2048 pes, 415 cpu-hours	Post-bounce evolution of 11 solar mass star, s11.2, 100 t steps	weak(k=8)

application software	application metric	target platform	problem	joule result
(07)GTC-S	Number of particles / (compute time / physical time step)	**Cray XT4** 64, 4096 pes 575 cpu-hours	PIC microturbulence plasma study in DIIID tokamak exp shot 122338 at 1.6s T, rho profile	weak(k=64)
(07)S3D INCITE	Compute time / dof / physical time step / processor core	**Cray XT3 U XT4** 14112 pes 41800 cpu-hours	Premixed methane-air, slot bunsen; non-premixed ethylene-air, planar slot jet	weak(k=1.96)
(06)DCA/QMC	Compute time / Green function update/ time slice	**Cray X1E** 512 pes 6352 cpu-hours	Pairing interaction study of 2d Hubbard Model	performance 74.03%
(06)ENZO	Compute time / processor core / physical time step	**IBM Power5** 512 pes 6725 cpu-hours	AMR (4lev) study, High red shift galaxy formation, 512^3 grid, 512^3 dark matter particles	performance 66.5%

application software	application metric	target platform	problem	joule result
(06)MADNESS	**time for** projection, compression, reconstruction, multiplication, differentiation	**Cray XT3** 4096 pes 7430 cpu- hours	Project nuclear potential from 4096 Cu atoms bcc lattice into wavelet basis w/ 1.e-3 precision	performance 77.27%
(06)ScalaBLAST	Compute time / query / processor core	**HP Itanium-2 (LP)** 1500 pes 45357 cpu-hours	Whole genome sequencing of Sargasso Sea environmental samples vs nr protein data base	new result, sequenced 1.2 million previously unknown proteins
(05)AORSA	compute time of FFT, ax=b	**Cray X1E** 256 pes 533 cpu-hours	Absorption of rf power by non-Maxwellian bulk ion components in NSTX tokamak	performance 55.75%
(05)CCSM	Simulated years / wall clock day	**Cray X1E** ~ 11904 cpu-hours	CAM3; spectral Eulerian dynamical core study (semi-Lagrangian vs Finite Volume	performance 53.7%

application software	application metric	target platform	problem	joule result
(05)LAMMPS	Dominated by force computation ala classical pairwise interactions	**IBM BG/L**	Md simulation of metal island on metal or oxide substrate to study effects of stress on device performance	new result,improved potentials (Yukawa, Morse, Buckingham); resolved dependence of stress in island in island size and adhesion to substrate
(05)Omega3P	Compute time / eigenmode / processor core	**IBM SP Power3** 768 pes 1753 cpu-hours	HOM study of 9-cell superconducting accelerating cavity in the ILC Tesla Test Facility	performance 81.3%
(05)S3D INCITE	Compute time / grid point / physical time step	**Cray X1** 256 pe	Non-premixed CO/ H2/N2-air plane jet flame simulation	performance 57.74%
(05)S3D SciDAC	Compute time / grid point / physical time step	**HP Itanium-2 (LP)** 256 pes metric only	Fuel spray injection study of effects of droplet size on evolution of carrier gas field features	performance 75%

application software	application metric	target platform	problem	joule result
(04)CCSM	Simulated years / wall clock day	**IBM p690**	T42(2.8d) T85	(Q2) 5 sim yrs / wall clock day, (Q4) > 38 sim yrs / wcd
(04)MILC	Compute time / sparse linear system; compute time / SU(3) matrix vector product	**QCDOC**	Single mass CG inverter on 128 QCDOC nodes	performance 90%
(04)NSM MC	Compute time / nucleon / shell / sample / imag time step	**IBM SP Power3**	Mo92, gds / 65536 samples	298MFlops, 74hours, 2048 pes
(04)RMPS	Compute time / Ax=kx solve / pe	**IBM SP Power3**	Electron impact excitation in DIIID tokamak energy and particle confinement study	Larger inversions , heavier atomic systems (235 level,Ne)
(04)VH-1	Compute time / zone update / processor core	**Cray X1**	3D SASI, l=1 mode	1,140,000 zone updates / second / pe

The Joule Software Metric
SC GG 3.1/2.5.2 Improve Computational Science Capabilities
FY09 Activity Report

Application Credits

VisIt

Sean Ahern (Oak Ridge National Laboratory, Oak Ridge, TN)
URL: http://www.llnl.gov/visit/

RAPTOR

Joseph C. Oefelein (Sandia National Laboratories, Livermore, CA)
URL: http://public.ca.sandia.gov/crf/research/index.php

XGC1

Choong-Seock Chang (Courant Institute of Mathematical Sciences, New York University, NY, NY)
URL: http://w3.physics.lehigh.edu/~xgc/, www.cims.nyu.edu/cpes/

CAM

James Hack (Oak Ridge National Laboratory, Oak Ridge, TN)
URL: http://www.ccsm.ucar.edu/models/atm-cam/

Additional Credits

Kenneth Roche, Ricky Kendall, Doug Kothe (ORNL)

DOE Program Contacts

Christine Chalk (christine.chalk@science.doe.gov)
Barbara Helland (helland@ascr.doe.gov)
Daniel Hitchcock (daniel.hitchcock@science.doe.gov)
Michael Strayer (michael.strayer@science.doe.gov)

Additional Contacts

Doug Kothe (kothe@ornl.gov)
Kenneth Roche (rochekj@ornl.gov)

Technical Team

(VisIt) Dave Pugmire, Tom Evans (ORNL), Hank Childs (LLNL); *(RAPTOR)* Ramanan Sankaran (ORNL); *(XGC1)* Scott Klasky, Pat Worley, Ed D'Azevedo (ORNL), Seung-Hoe Ku (Courant Institute of Mathematical Sciences, New York University), Mark Adams (Columbia University) *; (CAM)* Jim Rosinski, Pat Worley, Kate Evans (ORNL)

Target Machine : Cray XT5 at NCCS during FY09

QuadCore AMD Opteron (TM)	2.3e9 Hz clock	4 FP_OPs / cycle / core 128 bit registers
PEs	19,200 nodes	153,600 (149,504) cpu-cores (processors)
Memory	• 16 GB / node • 2 MB shared L3 / chip • 512 KB L2 / core • 64 KB D,I L1 / core	dual socket nodes 800 MHz DDR2 DIMM 3.2 GBps / core memory bw (25.6)
Network	AMD HT SeaStar2+	3D torus topology 6 switch ports / SeaStar2+ chip 9.6 GBps interconnect bw / port
Operating Systems	variant of Linux (xt-os2.1.50HD)	SuSE Linux on service / io nodes

Observation : a terascale / petascale supercomputer

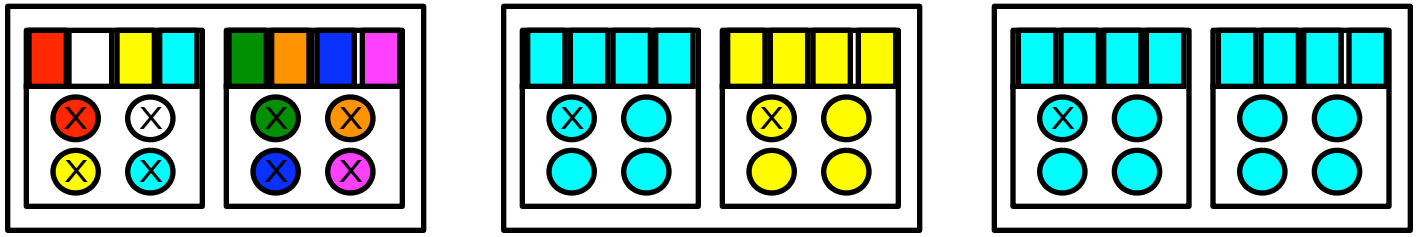
- Aggregated Cycle rate :
  2.3e9 cycles / second / cpu-core * 8 cpu-core / node * 18,688 nodes ~ 343.85 THz
- Aggregated Memory :
  18,688 nodes * 16 * 2^30 BYTES ~ 321.057 TB
- Peak FLOP rate :
  343.8592 THz * 4 FP_OP / cycle ~ 1.375 PFLOPs !

NUMA Node Structure of XT5 --> Hybrid Programming Model

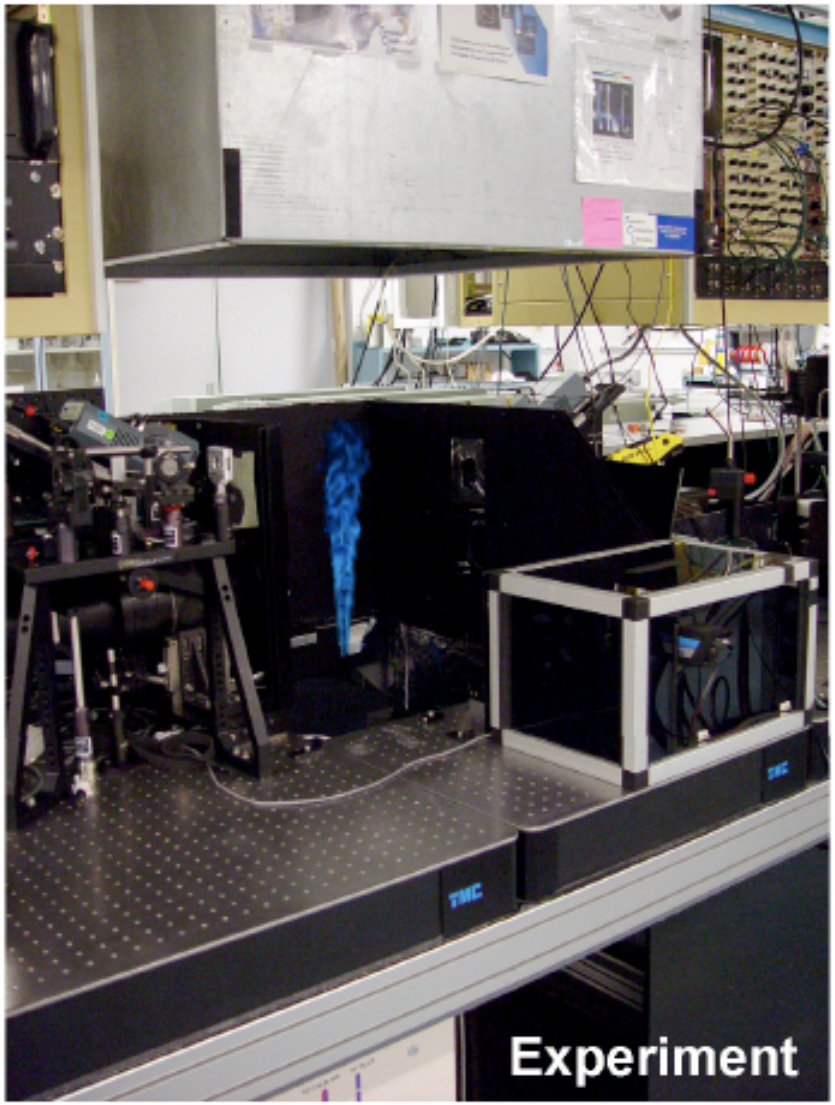
- MPI processes spawn lightweight processes
- OpenMP threads, #include <omp.h> , omp_set_num_threads();
- POSIX threads, #include <pthread.h> , pthread_create();

-lsize=8	MPI	LWP	DRAM
aprun -n <1-8>	1 - 8	1	2 * 2^30
aprun -n 2 -sn 2 -S 1 -d 4	2	1 - 4	4 * 2^30
aprun -n 1 -N 1 -d 8	1	1 - 8	16 * 2^30

<-S> * <-d> cannot exceed the maximum number of CPUs per NUMA node



Raptor



1. study the effects of LES grid resolution on scalar-mixing processes

2. understand the relationship between the grid spacing and the measured turbulence length scales from a companion set of experimental data (DLR-A, shown here)

3. study the effects of increasing jet Reynolds number on the dynamics of turbulent scalar-mixing

DLR-A Flame: Re_d = 15,200

Fuel: 22.1% CH_4, 33.2% H_2, 44.7% N_2

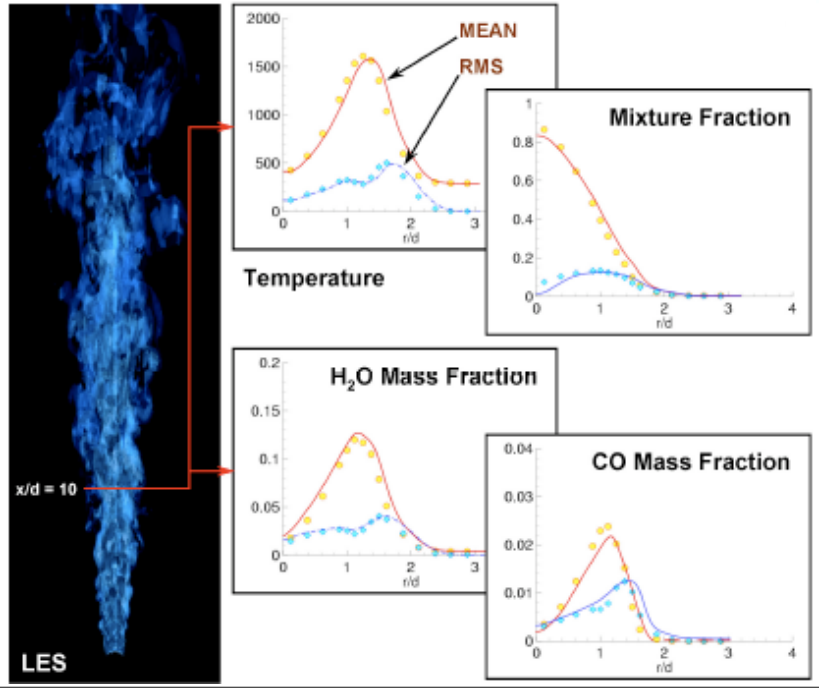
Coflow: 99.2% Air, 0.8% H_2O

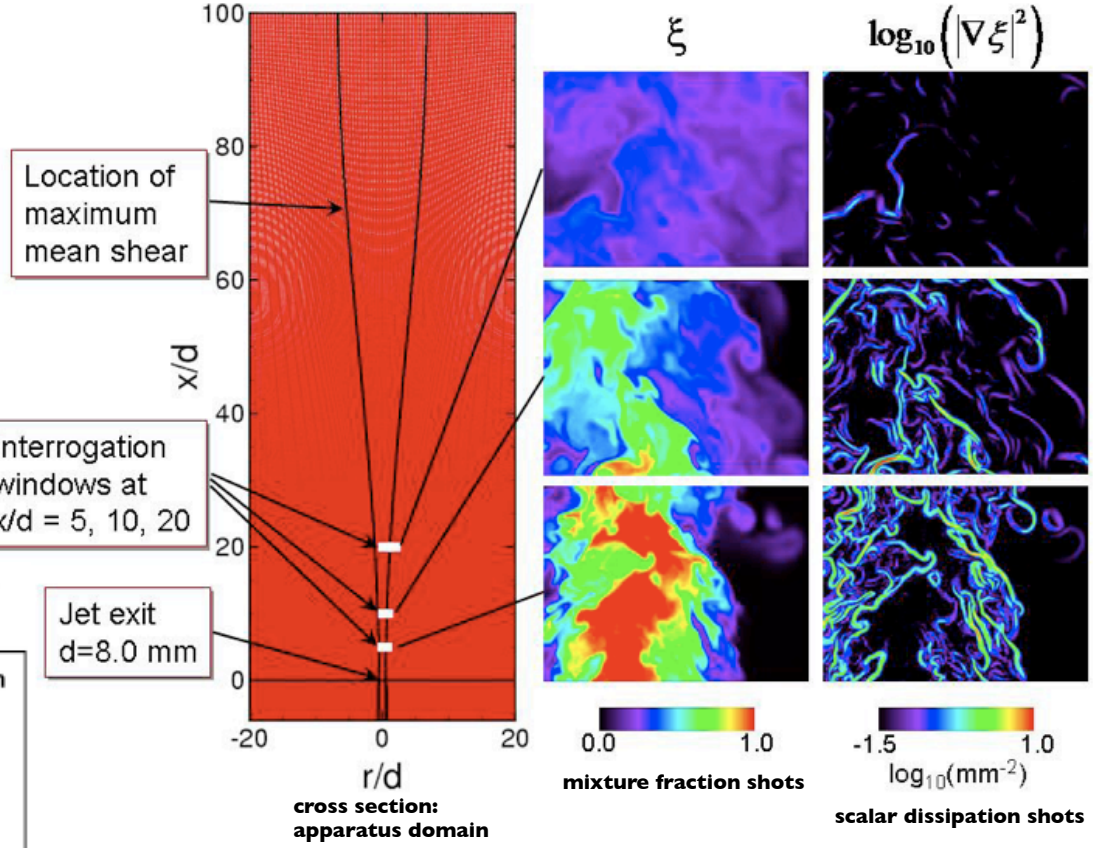
Detailed Chemistry and Transport: 12-Step Mechanism (J.-Y. Chen, UC Berkeley)

Raptor

Grid Number	Total Cells	Δt (Re_d = 15,200)
1	1,285,632	1.00 μs
2	10,285,056	0.50 μs
3	82,280,448	0.25 μs

50 physical time steps per grid





Domain: entire burner geometry (inside the jet nozzle and the outer co-flow) + downstream space around burner
Inner nozzle diameter : 8.0 mm
Outer nozzle : surface is tapered to a sharp edge at the burner exit
Specifics: 110 inner jet diameters in the axial direction (88cm) x 40 jet diameters in the radial direction (32 cm)

Raptor

METRIC : **CPU time / Number of Grid cells / Number of time-steps**

(1, 034 seconds x 47, 616 cores) / 10, 285, 056 cells / 50 time-steps = 0.096

(It cost 96-milliseconds of processor time per cell per time-step to simulate the problem on 47,616 cores.)

PROBLEM	PEs	Total Instructions	Floating Point INS	Wall Time(s)
DLR-A: grid 2, 50 timesteps	47,616	2.06E+17	3.78E+14	1425.761

* Difference in wall and cpu time: the initialization step when the computational mesh and initial condition information were read from the disk

CAM

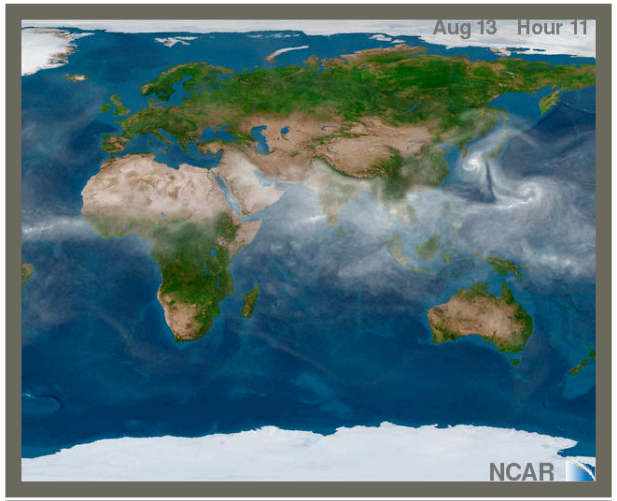
CCSM Community Atmospheric General Circulation Model for Weather and Climate Research

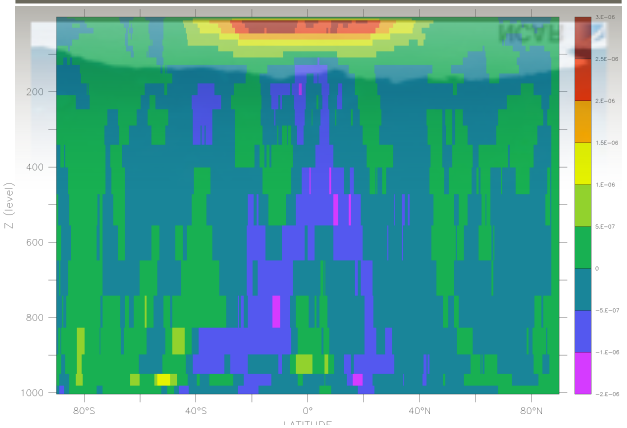
- Physical parameterizations for prognostic cloud moisture, radiative effect of aerosols, long, short wave radiation interaction, interfaces with land & ocean

- Phases:
  - Dynamics: evolution equations for atmospheric flow
    - spectral Eulerian
    - spectral semi-Lagrangian
    - finite-volume semi-Lagrangian
  - Physics: subgrid-scale phenomena such as precipitation processes, clouds, long and short wave radiation transfer, and turbulent mixing

- Physical parameterizations only in vertical dimension, all on-processor, and not load balanced

- Is the current performance bottleneck to CCSM





- Differences in short-wavelength heating rates between 2 T341 CAM runs w/ & w/o volcanic aerosols
- Oct 1991 average during Mount Pinatubo eruption
- Y axis depicts vertical pressures (mbars)
- Red signifies areas where the volcanic CAM run has more heating

CAM

- T341 horizontal resolution
  (1024 latitude x 512 longitude) with 26 vertical levels

- Execute in uncoupled (standalone) mode with fully active Common Land Model
  Sea surface temperatures & sea ice concentrations via external forcing datasets

- Spectral Eulerian dynamic core within CAM3.5

- Improve model run time for a one-month simulation.
  One month adequately represents run-time contributions from all components

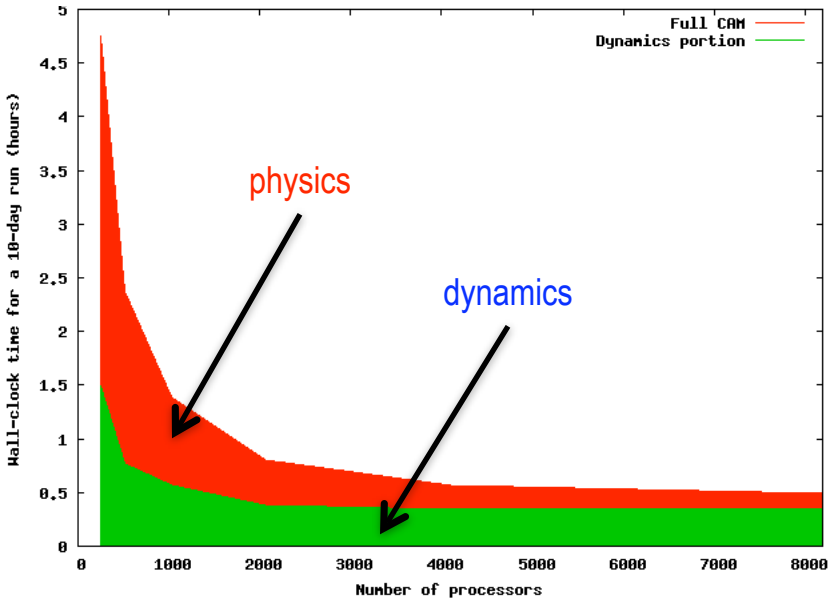
- Constant time step of 150 seconds
  integrate 17856 time steps

Performance: focus on OpenMP improvements to the decomposition of the spectral Eulerian dynamical core

CAM

- Dynamics percentage increases with core count
- 8192 PEs
- Most CAM configurations have plenty of headroom on node memory
- Single-threaded I/O, limited exploitation of potential parallelism in spectral Eulerian dynamics, inherent load imbalance in physics limits scaling
- Communication is a small part of the total computation cost

CAM strong scaling on Jaguar/XT5 with T341 mesh



Performance Data	Atmosphere	CLM	I/O	Total
Time (s)	5916.475	112.048	115.024	6481.724
FP Instructions	2.13×10^{15}	3.89×10^{13}		2.17×10^{15}

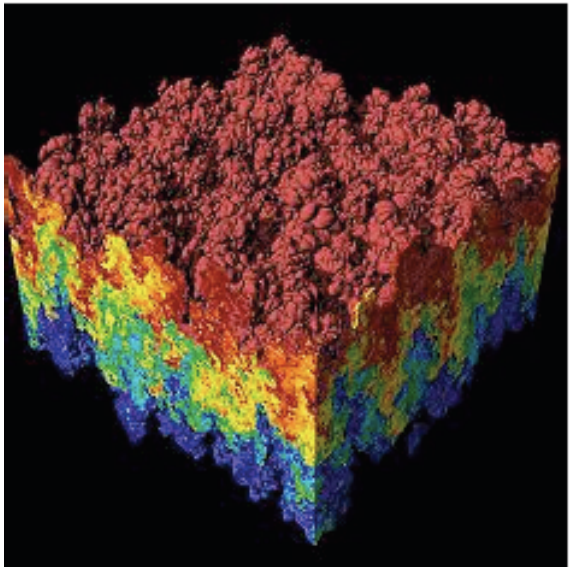
~1 simulated year
/ 22 hours wall-clock

VisIt

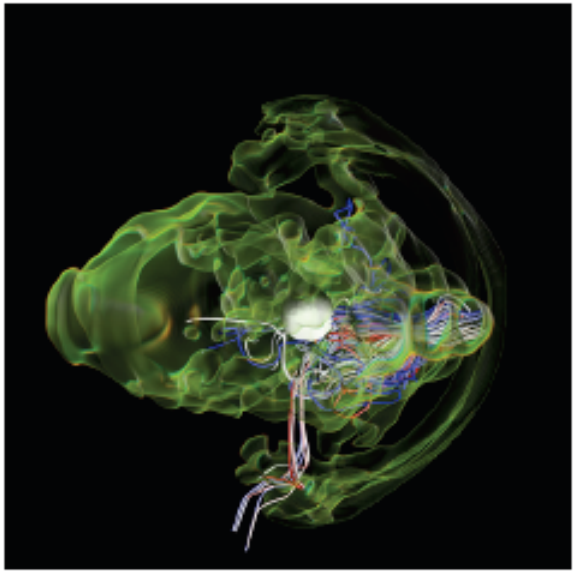
data exploration quantitative analysis comparative analysis

visual debugging communication of results remote visualization

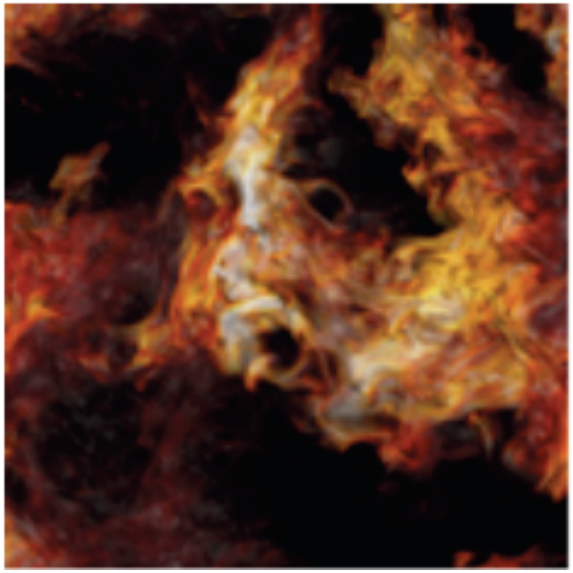
Isosurface Extraction



Streamlining



Volume Rendering



* large data strategy is to use distributed memory data parallelism

* each process creates an identical data flow network and works in MIMD model

VisIt

Denovo: Study the radiation dose concentrations around a reactor core in a nuclear power generating plant

steady state Boltzmann transport calculation

4096 spatial domains

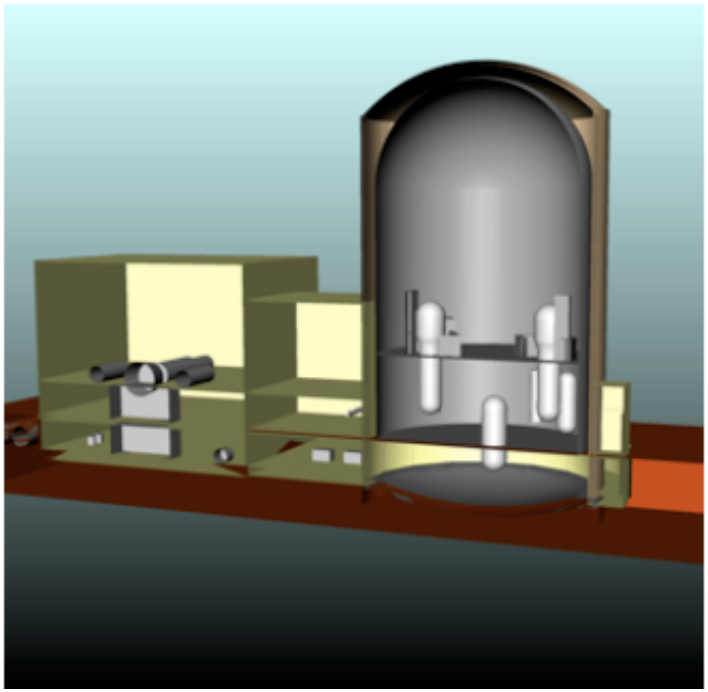
materials: concrete, reactor fuel, steel, reduced density steel, air

mesh size : 456 x 648 x 351

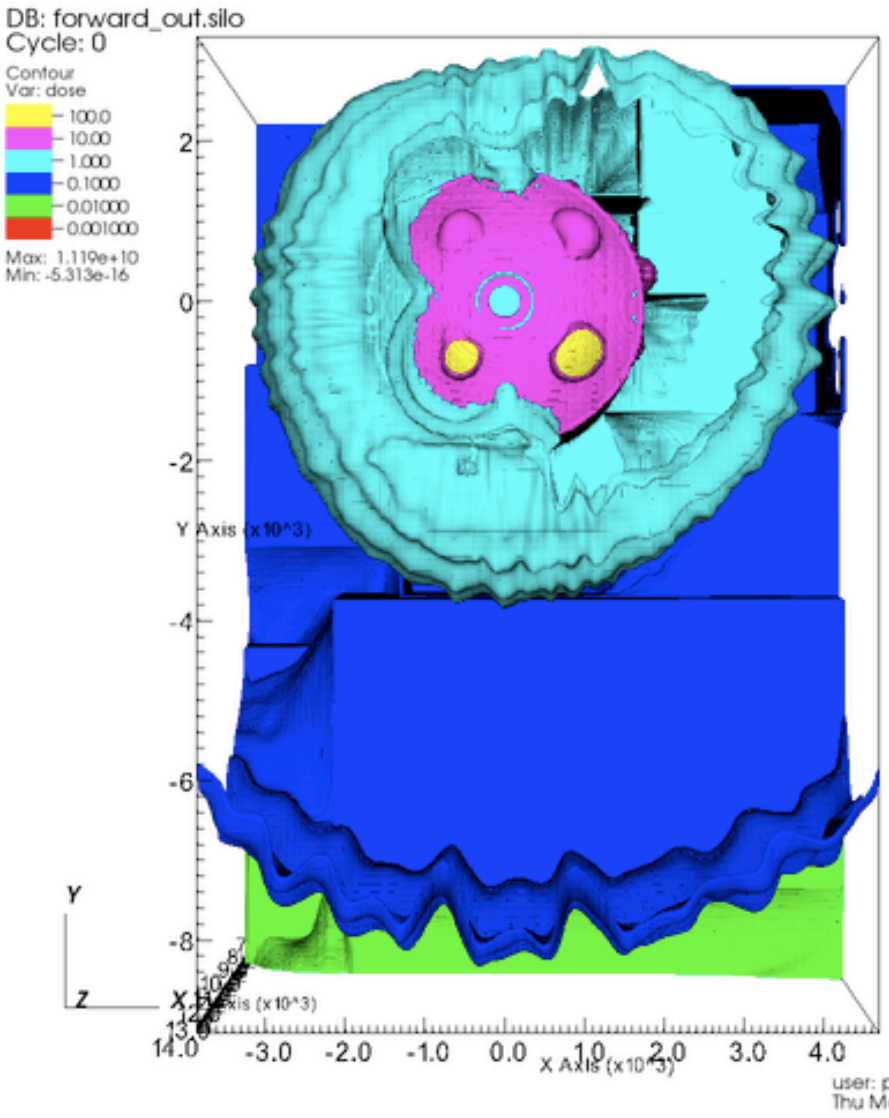
scalar flux values for 27 energy groups per zone

4096 files (1 per domain) totaling 36.23843 GB

-double precision values



VisIt



Isosurface extraction :

extract the three dimensional points in a volume with a specific value and connect them with a continuous surface

Contours at isovalues :
0.001, 0.01, 0.1, 1.0, 10.0, 100.0

Resolution : 1024 x 1024 pixels

***27 energy level flux values used by VisIt to compute the dose variable scalar field**

VisIt : Q2 Performance Results

METRIC : time to render a frame

Isosurface (per core times)	Minimum	Maximum	Average
Pipeline			
Isosurface	0.0140	0.0270	0.01768
Render	0.020	0.065	0.02245
Scalable Rendering	0.048	0.087	0.05193
Expression Engine	0.181	0.245	0.21097

Volume Render (2000_{samp})	Minimum	Maximum	Average
Pipeline	28.911	29.018	28.92484
Volume Render	28.716	28.78	28.7293
Sample Point Extraction	0.1	0.332	0.25329
Sample Point Comm.	28.335	28.465	28.41073
Image Communication	0.00000	0.215	0.00741
Expression Engine	0.156	0.199	0.16275

PROBLEM	PEs	Total Instructions	Floating Point INS	Wall Time(s)
Isosurface	4096	9.47828E+14	1.73459E+11	247.499
Volume Render (2000)	4096	1.03872E+15	1.78848E+11	194.511

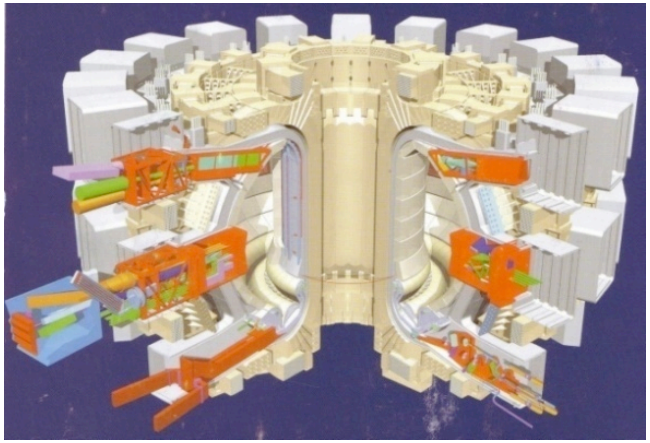
* only the 2000 sample volume rendering of the Denovo data is reported here. to see the other benchmark numbers, reference the Q2 report.

XGC1

5D Gyrokinetic Full-Function Particle-in-Cell Model for Whole Plasma Dynamics in Experimentally Realistic Magnetic Fusion Devices

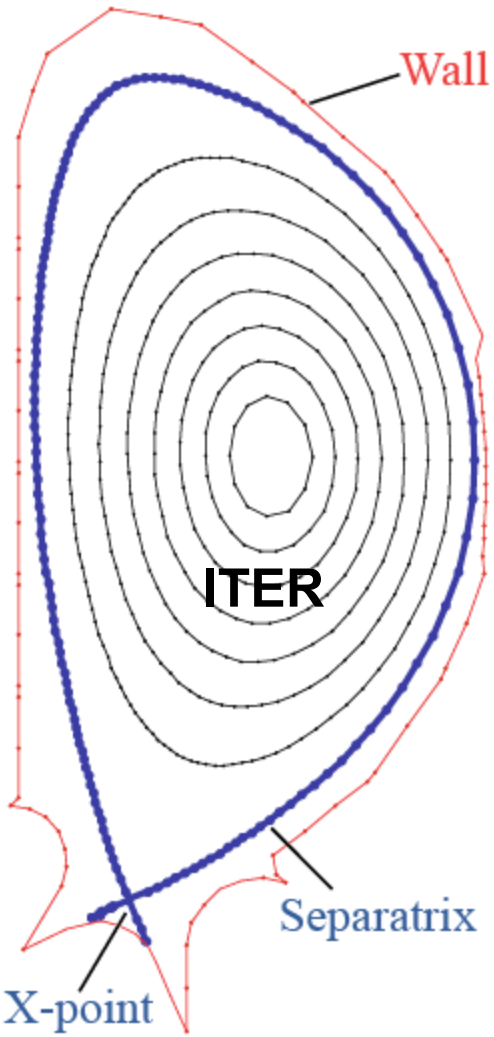
$$\frac{df}{dt} = [\frac{\partial}{\partial t} + (v_{\|}\mathbf{b} + \mathbf{v}_d + \mathbf{v}_{E\times B}) \times \frac{\partial}{\partial \mathbf{R}} - \mathbf{b}^* \cdot \nabla(\mu B + < \Phi >_{\alpha}) \frac{\partial}{\partial v_{\|}}] f$$

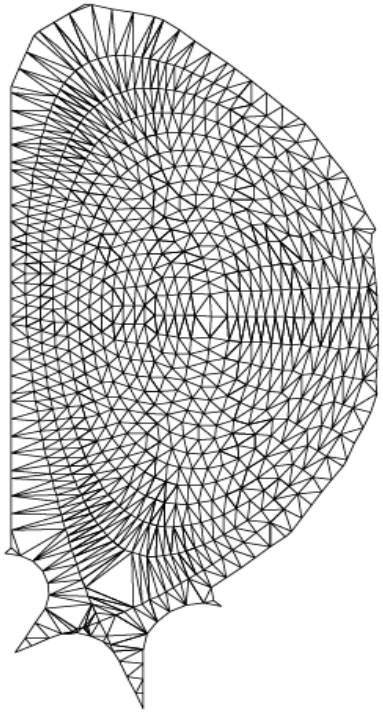
$$\mathbf{B}^* = \mathbf{B} + (B v_{\|}/\Omega_s)\nabla \times \mathbf{b} \qquad \mathbf{b}=\mathbf{B}/B \qquad \text{(Electrostatic eqn.)}$$



- Gyrokinetic “full-f” PIC model of magnetic fusion plasmas, with inclusion of magnetic separatrix, magnetic X-point, conducting material wall, & momentum/energy conserving Coulomb collisions
- Full-f description allows turbulence and background plasma to interact self-consistently and background plasma to evolve to a self-organized state
- Focus: understand and predict plasma transport and profile in the “edge pedestal” around separatrix

XGC1





Unstructured triangular mesh on numerical B data

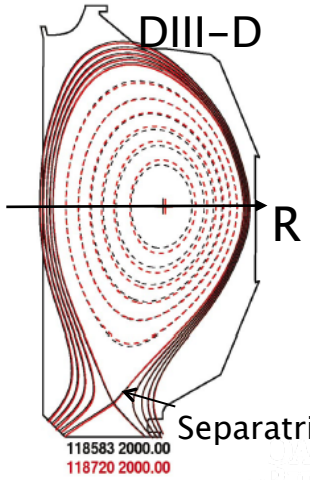
→ Extra difficulty in particle sorting and interpolation

- Fixed unstructured grid following equilibrium magnetic field lines with embedded discrete marker particles representing ions, electrons, and neutral particles
- Marker particles time-advanced with Lagrangian equation of motion (either 4th order PC or 2nd order RK)
- Marker particle charges accumulated on grid, followed by gyrokinetic Poisson solve for electrostatic field
- PETSc for Poisson solve, ADIOS for I/O, Kepler for workflow, Dashboard for monitoring/steering

XGC1

Performance Metric	Value
Cores	29,952
Cycles per second per core	2279.67×10^6
Instructions per second per core	2474.96×10^6
Floating-point operations per second per core	223.63×10^6
Particle time steps processed per second	0.639×10^9

- DIII-D tokamak at General Dynamics with realistic physical and diverter geometry including material wall
- 13.5B particles; 1 MPI process per core
- 29,952 processor cores on Jaguar/XT5 for a full 24-hour simulation (16.5 hours of data collection)



- Performance: rate at which particles can be integrated forward in time
- Q2 result: DIII-D tokamak (13.5B particles on 29,952 processor cores)
- Q4 goal: ITER tokamak (5X Q2 problem size: 67.5B particles on 149,760 processor cores)

Application	CCSM: CAM	RAPTOR	VisIt Isosurface	XGC1
Problem	•Simulated month •T341 mesh •Constant 150 second time step	•DLR-A configuration •50 time steps •10,285,056 cell mesh	•1024^2 pi •.001,.01,.1,1, 10,100 Nuclear Reactor •456 x 648 x 351 •4096 core •27 groups	DIII-D 0.6B particles/s
Metric	Simulated years / wall clock day	•0.096 grind time •Time / Number of grid cells /Number of time steps	Time to render a frame	•Particles processed per second
PEs	8,192	47,616	4,096	29,952
Time [s]	6,481.72	1,425.76	247.49	52,766.01
Instructions		2.06E+17	9.48E+14	3.91E+18
Floating Point	2.17E+15	3.73E+14	1.73E+11	3.53E+17

Preview of Q4 Results

- **RAPTOR halo-exchanges are nearest neighbor only**
  - Initial configuration … send/receive calls in pairs corresponding to each neighbor
  - Receive calls can be posted early as long as the buffer is available
  - No need to wait on 'send' calls until the send buffer is about to be altered
  - Interleave computation to give more breathing room for communication

- **METRIC : CPU time / Number of Grid cells / Number of time-steps**

  Q2: (1033.83 seconds x 47,616 cores)/(10,285,056 cells)/(50 time-steps) = 0.0957

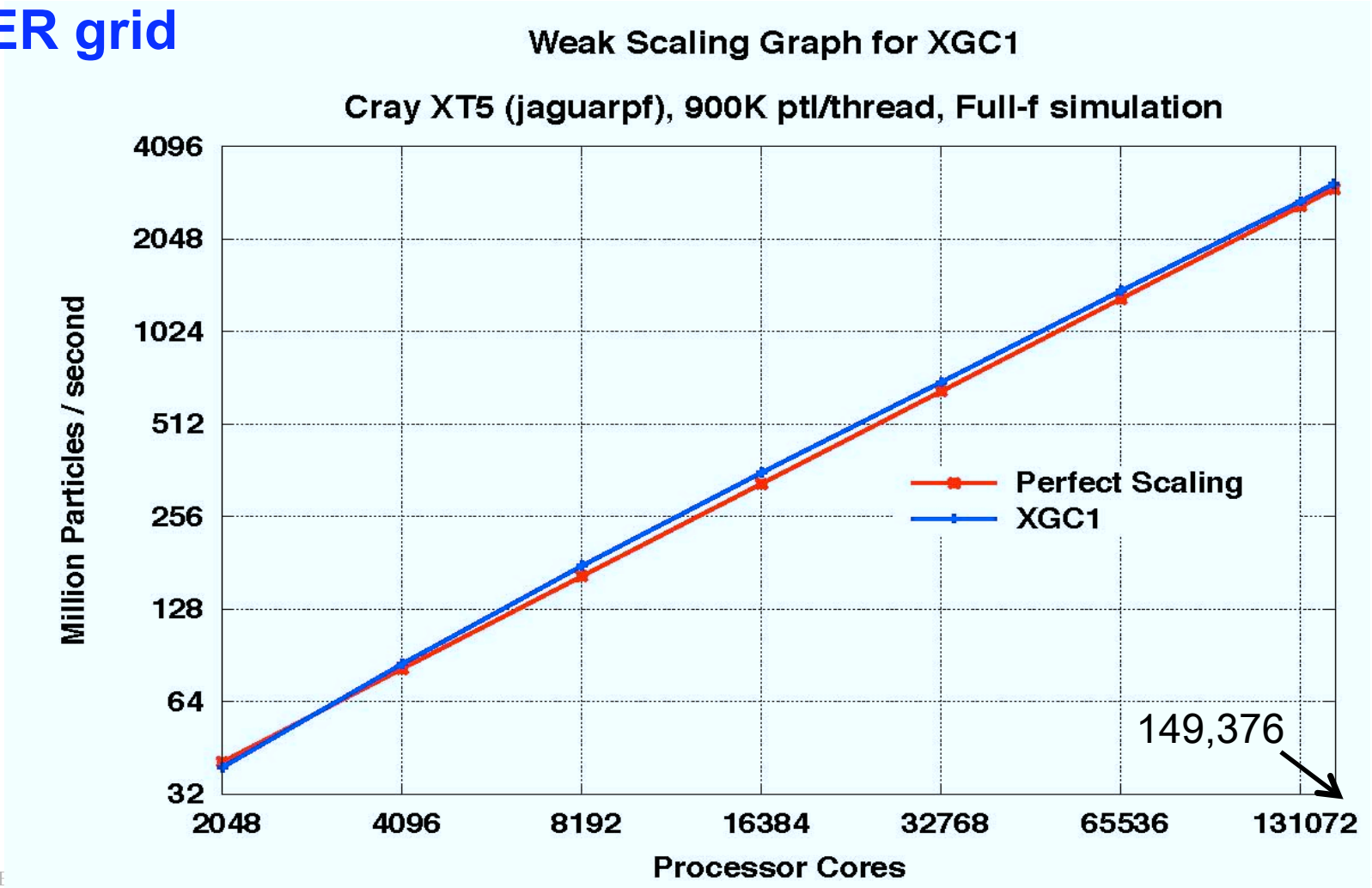
  Q4: (750.960 seconds x 112,320 cores)/(24,261,120 cells)/(50 time-steps) = 0.0411

Problem	Cores	Total Instructions	Floating Point Instructions	Wall Time, Seconds
Q2 (DLR-A): 10,285,056 cells	47,616	2.06E+17	3.78E+14	1034
Q4 (DLR-A): 24,261,120 cells	112,320	4.86E+17	8.92E+14	444

Preview of Q4 Results

MPI + OpenMP

on ITER grid



Example Summary of Results from the FY08 Benchmarks:

Application	DCA++	GYRO	PFLOTRAN
Metric	time / disorder configuration	timesteps / second / process	time / dof / PE
Problem	$N_{dis} = 64, N_c = 16, N_t = 150$	$\mu = 30$, 10 timesteps	$64.8M$ DOFs, 200 flow, transport steps
Hardware Used	7808 PEs	4608 PEs	4000 PEs
Walltime	25339 s	17.23 s	2594 s
Instructions	5.1805×10^{17}	2.2410×10^{14}	2.2222×10^{16}
Floating Point Ops	4.6270×10^{17}	6.8320×10^{13}	1.2898×10^{15}

Application	DCA++	GYRO	PFLOTRAN
Metric	time / disorder configuration	timesteps / second / process	time / dof / PE
Problem	$N_{dis} = 256, N_c = 16, N_t = 150,$	$\mu = 40$, 10 timesteps	129, 635, 520 DOFs, Q2 stepping
Hardware Used	31232 PEs	24576 PEs	8000 PEs
Walltime	23791 s	152.75 s	2958.36 s
Instructions	1.9300×10^{18}	1.2202×10^{16}	5.0374×10^{16}
Floating Point Ops	1.8126×10^{18}	6.0882×10^{15}	2.8603×10^{15}

TOTALS	Q2	Q4	ratio (Q4 : Q2)
$\sum$ Walltime	27950.23 s	26902.11 s	.9625
$\sum$ PEs	16416	63808	3.8869
$\sum$ Instructions	5.4049×10^{17}	1.9925×10^{18}	3.6866
$\sum$ Floating Point Ops	4.6405×10^{17}	1.8215×10^{18}	3.9253

Thank you.

Kenneth J. Roche, Future Technologies Group
Computer Science and Mathematics Division
Oak Ridge National Laboratory
rochekj@ornl.gov

Douglas B Kothe, Science Director
National Center for Computational Sciences
Oak Ridge National Laboratory
kothe@ornl.gov